Apache Hive docker images.

Apache Hive

Apache Hive™ is a popular open-source data warehousing framework that allows users to query large datasets stored in distributed storage systems like Hadoop HDFS, Apache Ozone, Amazon S3, or Microsoft Azure Data Lake Storage (ADLS). It provides a SQL-like (called HiveQL) interface to query, analyze, and manage structured and semi-structured data.

With Hive, users can define tables and schemas, write queries in a familiar SQL-like language and perform various data analysis tasks such as filtering, aggregating, and joining data. Hive supports many popular data formats including CSV, JSON, Avro, ORC, and Parquet.

Quickstart

export HIVE\_VERSION=<Hive Tag/Image Version>

Example:

export HIVE\_VERSION=4.0.1

Launch the HiveServer2 with an embedded Metastore.

This is lightweight and for a quick setup, it uses Derby as metastore db.

docker run -d -p 10000:10000 -p 10002:10002 --env SERVICE\_NAME=hiveserver2 --name hive4 apache/hive:${HIVE\_VERSION}

Launch Standalone Metastore

For a quick start, launch the Metastore with Derby,

docker run -d -p 9083:9083 --env SERVICE\_NAME=metastore --name metastore-standalone apache/hive:${HIVE\_VERSION}

Warning! Everything would be lost when the service is down! In order to save the Hive table's schema and data, start the container with an external Postgres and Volume to keep them.

Advanced Setup

Using HS2 with Standalone/Remote Metastore

docker run -d -p 10000:10000 -p 10002:10002 --env SERVICE\_NAME=hiveserver2 \

--env SERVICE\_OPTS="-Dhive.metastore.uris=thrift://metastore:9083" \

--env IS\_RESUME="true" \

--name hiveserver2-standalone apache/hive:${HIVE\_VERSION}

"-Dhive.metastore.uris is used to specify the external Metastore

Launch Standalone Metastore With External RDBMS (Postgres/Oracle/MySql/MsSql)

docker run -d -p 9083:9083 --env SERVICE\_NAME=metastore --env DB\_DRIVER=postgres \

--env SERVICE\_OPTS="-Djavax.jdo.option.ConnectionDriverName=org.postgresql.Driver -Djavax.jdo.option.ConnectionURL=jdbc:postgresql://postgres:5432/metastore\_db -Djavax.jdo.option.ConnectionUserName=hive -Djavax.jdo.option.ConnectionPassword=password" \

--mount source=warehouse,target=/opt/hive/data/warehouse \

--name metastore-standalone apache/hive:${HIVE\_VERSION}

Note:

To save the data between container restarts, you can start the HiveServer2 with mounted volume:

docker run -d -p 10000:10000 -p 10002:10002 --env SERVICE\_NAME=hiveserver2 \

--env SERVICE\_OPTS="-Dhive.metastore.uris=thrift://metastore:9083" \

--mount source=warehouse,target=/opt/hive/data/warehouse \

--env IS\_RESUME="true" \

--name hiveserver2 apache/hive:${HIVE\_VERSION}

Any other cases your data will not persist!

Custom Configurations

If you want to use your own core-site.xml/hdfs-site.xml/yarn-site.xml or hive-site.xml for the service, you can provide the environment variable HIVE\_CUSTOM\_CONF\_DIR for the command. For example:

Put the custom configuration file under the directory /opt/hive/conf and run:

docker run -d -p 9083:9083 --env SERVICE\_NAME=metastore \--env DB\_DRIVER=postgres -v /opt/hive/conf:/hive\_custom\_conf --env HIVE\_CUSTOM\_CONF\_DIR=/hive\_custom\_conf \--name metastore apache/hive:${HIVE\_VERSION}

Usage

Accessing Beeline:

docker exec -it hive4 beeline -u 'jdbc:hive2://localhost:10000/'

You can use beeline as it is installed on your host machine beeline -u 'jdbc:hive2://localhost:10000/'

Accessing HiveServer2 Web UI:

Accessed on browser at http://localhost:10002/

Example Queries to execute in Beeline

show tables;

create table hive\_example(a string, b int) partitioned by(c int);

alter table hive\_example add partition(c=1);

insert into hive\_example partition(c=1) values('a', 1), ('a', 2),('b',3);

select count(distinct a) from hive\_example;

select sum(b) from hive\_example;

Support

* Apache Hive Website: [https://hive.apache.org](https://hive.apache.org/)
* Apache Hive Github Repo: <https://github.com/apache/hive>
* Apache Hive Jira(To contribute or Report Fixes/Improvements): <https://issues.apache.org/jira/projects/HIVE/issues>
* Apache Hive Getting Started Wiki: <https://hive.apache.org/developement/gettingstarted>
* Apache Hive Mailing Lists(For any queries): <https://hive.apache.org/community/mailinglists>

Further Reading

<https://hive.apache.org/developement/quickstart/>